Project 4: Bank Marketing

02/28/2020

1 Introduction

1.1 Background

1.2 Statistical questions of interest

2 Analysis Plan

2.1 Population and study design

The dataset is downloaded from UCI Machine Learning Repository and is related to direct marketing campaigns of a Portuguese Banking institution. These campaigns were based on phone calls. Often, more than one calls were done to the same client to access if their product "term deposit" will be subscribed (yes) or not subscribed(no). This dataset is available at http://archive.ics.uci.edu/ml/datasets/Bank+Marketing. There were 4 datasets in it from which bank-additional-full.csv is used that has all examples (41188) and 20 inputs ordered by date (from May 2008 to November 2010). There are 20 input variables and 1 output variable (desired target). These dataset attributes denote customer data, socio-economic data, telemarketing data and some other data. Some attributes are numerical, and some are categorical. The dataset was loaded in R Studio and checked for any missing values using is.na function and found that it didn't have any missing values, so we have a clean dataset.

2.2 Descriptive Analysis

Table.1 Features description of the Bank Marketing Dataset (BMD).

Feature	Description	Attribution
`y`	Desired target. Has the client subscribed a term deposit?	binary
`age`	Client Age	numeric
`job`	Type of Job	categorical
`marital`	Client's marital status	categorical
`education`	Client's education	categorical
`default`	Has credit in default?	categorical
`housing`	Has housing loan?	categorical
`loan`	Has personal loan?	categorical
`contact`	Last contact month of year	categorical
`month`	Month of last contact with client	categorical
`day_of_week`	Last contact day of the week	categorical
`duration`	last contact duration, in seconds	numeric
`campaign`	Number of contacts performed during this campaign and for this client	categorical

Feature	Description	Attribution
`pdays`	Number of days that passed by after the client was last contacted from a previous campaign	numeric
`previous`	Number of client contacts performed before this campaign	numeric
`poutcome`	Outcome of the previous marketing campaign	categorical
`emp.var.rate`	Quarterly employment variation rate	numeric
`cons.price.idx`	Monthly consumer price index	numeric
`cons.conf.idx`	Monthly consumer confidence index	numeric
`euribor3m`	Daily euribor 3-month rate	numeric
`nr.employed`	Quarterly number of employees	numeric

By description, `duration` should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. So this attribute highly affects the output target. More than 95% values in `pdays` are 999, which means client was not previously contacted. Only 3 cases are `yes` in `default`, and more than 20% are unknown. So we remove these three variables.

3 Results

3.1 Descriptive Analysis

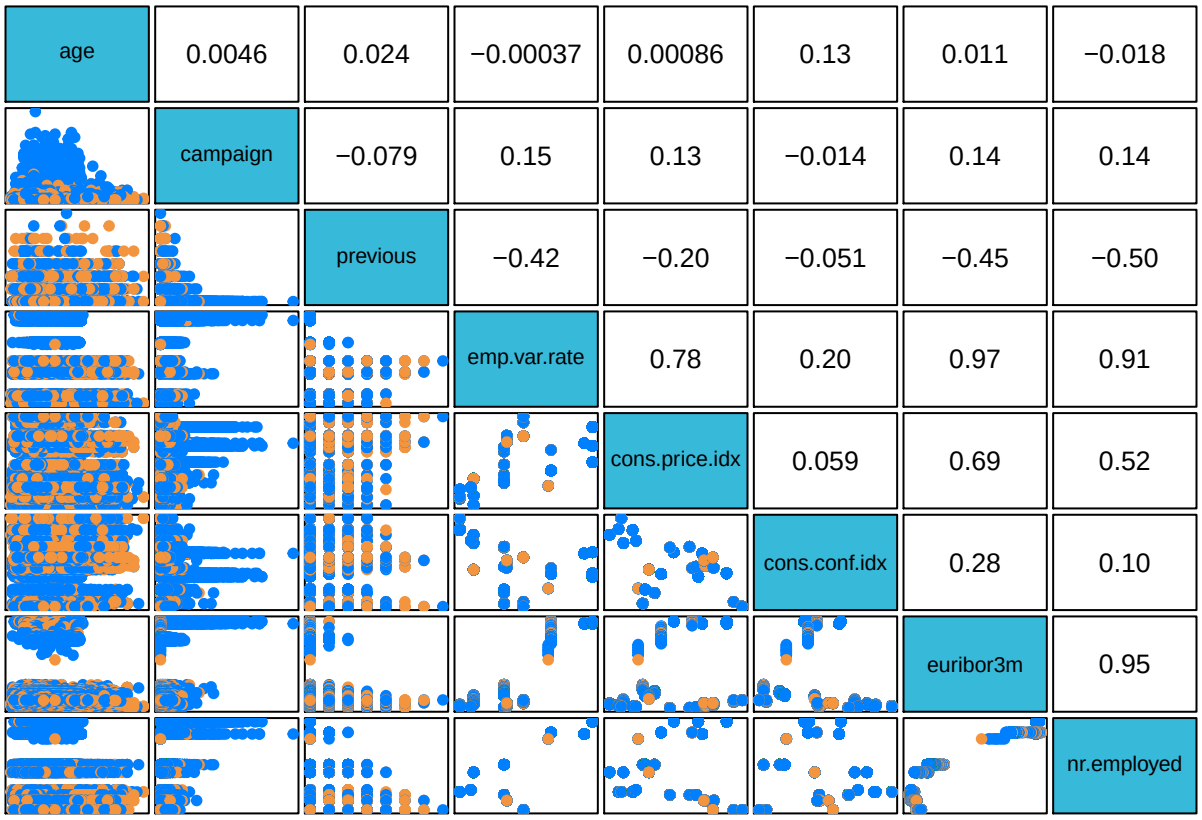

Figure.1 Scatterplot lower triangular matrix and correlation upper triangular matrix for all the quantitative features presented in the Bank Marketing Dataset (BMD).

In Figure 1 we see the scatterplots and correlations, two-by-two, for all the eight numerical features in the BMD. In more than half of them we see a random behaviour, that is also described by a correlation close to zero or between the interval -0.3 and 0.3. A (very) strong (and positive) correlation is seen in three cases. `emp.var.rate` vs. `euribor3m` (cor. 0.97), `euribor3m` vs. `nr.employed` (cor. 0.95), and `emp.var.rate` vs. `nr.employed` (cor. 0.91), i.e., involving only three features - employment variation rate, Euro Interbank Offered Rate (Euribor) and number of employees. To avoid high collinearlity, we only remain `nr.employed` to analysis.

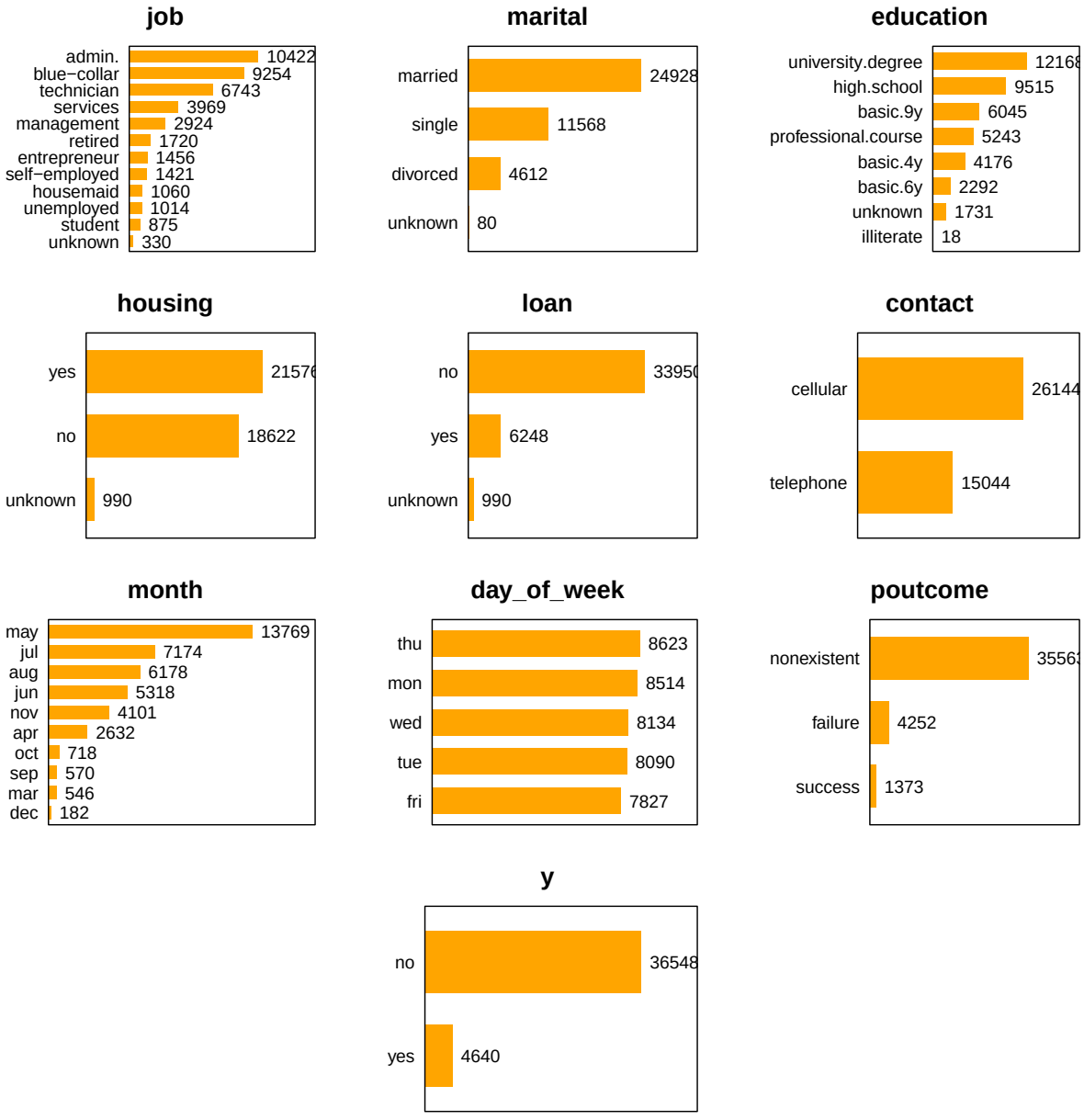

Figure.2 Bar plots for all the qualitative features presented in the Bank Marketing Dataset (BMD).

Already in Figure 2 we have the frequencies for each level of the categorical features in the BMD. First, we see that the desired target is unbalanced, with more than 85% of the observations corresponding to clients that didn't subscribed to a term deposit. An equilibrium between levels is only present in the `day.of.week`

last contact feature. By this Figure we can also see that the last contact of most of the clients was in may (`month` feature), that most of the clients have a nonexistent previous marketing campaign (`poutcome` feature), that they are married (`marital` feature) and that most have a `job` in the administrative sector.

3.2 Logistic Model

```
Call:
glm(formula = y ~ age + job + contact + month + day_of_week +
    campaign + poutcome + emp.var.rate + cons.price.idx + cons.conf.idx +
    nr.employed, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0480  -0.3882  -0.3201  -0.2593   2.9734

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.668e+02  2.892e+01  -9.227  < 2e-16 ***
age                 -3.299e-03  2.009e-03  -1.643 0.100455
jobblue-collar      -2.473e-01  5.918e-02  -4.180 2.92e-05 ***
jobentrepreneur     -5.688e-02  1.110e-01  -0.513 0.608279
jobhousemaid        -1.828e-01  1.326e-01  -1.379 0.167954
jobmanagement       -1.119e-02  7.829e-02  -0.143 0.886301
jobretired           2.251e-01  9.661e-02   2.330 0.019796 *
jobself-employed    -7.887e-02  1.063e-01  -0.742 0.458008
jobservices         -1.866e-01  7.560e-02  -2.468 0.013573 *
jobstudent           1.873e-01  1.004e-01   1.866 0.062104 .
jobtechnician       -2.536e-02  5.860e-02  -0.433 0.665159
jobunemployed       -8.838e-02  1.165e-01  -0.758 0.448250
jobunknown          -2.174e-01  2.199e-01  -0.988 0.323010
contacttelephone    -7.185e-01  7.051e-02 -10.189  < 2e-16 ***
monthaug             5.521e-01  1.119e-01   4.932 8.13e-07 ***
monthdec             5.978e-01  1.917e-01   3.118 0.001818 **
monthjul             9.804e-02  8.691e-02   1.128 0.259294
monthjun            -7.103e-01  1.165e-01  -6.099 1.07e-09 ***
monthmar             1.578e+00  1.315e-01  11.995  < 2e-16 ***
monthmay            -3.883e-01  7.429e-02  -5.227 1.72e-07 ***
monthnov            -3.093e-01  8.834e-02  -3.501 0.000464 ***
monthoct             1.509e-01  1.185e-01   1.273 0.202952
monthsep             3.744e-01  1.483e-01   2.524 0.011602 *
day_of_weekmon      -1.826e-01  6.100e-02  -2.993 0.002767 **
day_of_weekthu       9.846e-02  5.884e-02   1.673 0.094242 .
day_of_weektue       5.744e-02  6.081e-02   0.944 0.344919
day_of_weekwed       1.646e-01  6.027e-02   2.730 0.006325 **
campaign            -4.546e-02  9.780e-03  -4.648 3.35e-06 ***
poutcomenonexistent  4.867e-01  5.805e-02   8.384  < 2e-16 ***
poutcomesuccess      1.767e+00  8.283e-02  21.335  < 2e-16 ***
emp.var.rate        -1.492e+00  1.305e-01 -11.435  < 2e-16 ***
cons.price.idx       2.286e+00  2.079e-01  10.994  < 2e-16 ***
cons.conf.idx        3.945e-02  5.163e-03   7.641 2.16e-14 ***
nr.employed          1.013e-02  1.893e-03   5.349 8.85e-08 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25929  on 37068  degrees of freedom
Residual deviance: 20343  on 37035  degrees of freedom
AIC: 20411

Number of Fisher Scoring iterations: 6
```

3.3 Model Diagnostics

4 Discussion

5 Appendix. Reference

[1] [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

6 Appendix II. Group Partners

This Document is the project 3 of Team 7 in STA 207, Winter 2020.

1. Bingdao Chen bdchen@ucdavis.edu contribution: descriptive analysis and model establishment
2. Yahui Li yhuli@ucdavis.edu contribution: casual inference and conclusion
3. Zihan Wang zihwang@ucdavis.edu contribution: fixed effects model
4. Jian Shi jnshi@ucdavis.edu contribution: model diagnose

The repository in Github is on https://github.com/yhli097/STA207Project4.git